Abstract

A method and system identifying the language of a textual passage is disclosed. The method and system includes parsing the textual passage into n-grams and assigning an initial weight to each n-gram, and adjusting the weight initially assigned to a word or n-gram parsed from the textual passage. The initially assigned weight is adjusted in a manner proportionate to the inverse of the number of languages within which such words or n-grams appear. Reducing the weight assigned to such words or n-grams diminishes – without completely eliminating – their importance in comparison to other words or n-grams parsed from the same textual passage when determining the language of a passage. The method and system of the present invention appropriately weighs the short words or n-grams common to multiple languages without affecting the short words or n-grams that are uncommon to several languages.